

Yapay Öğrenme ile Biyolojik Dizilerin Anlamlandırılması

Volkan Atalay

ODTÜ

Bilgisayar Mühendisliği Bölümü

ve

Kanser Sistem Biyolojisi Laboratuvarı

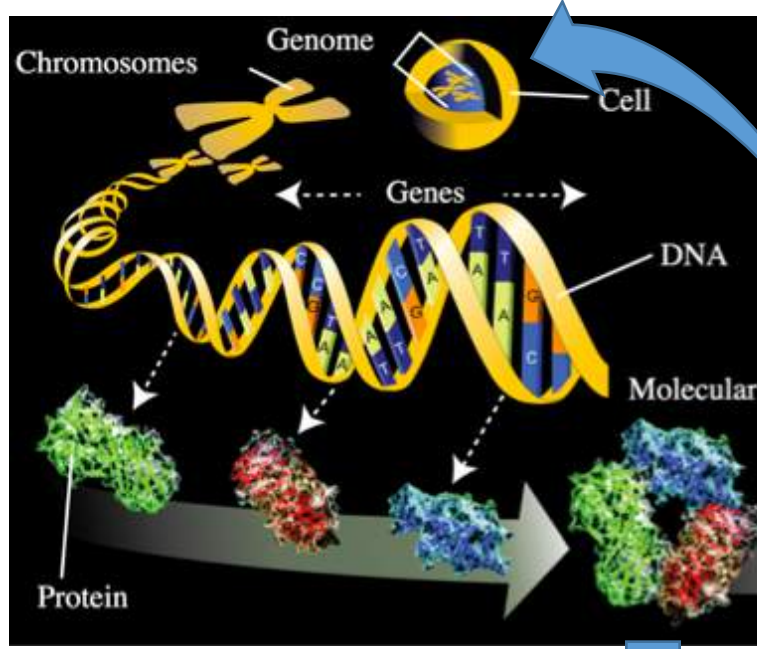
Giriş-Temel kavramlar

- DNA (kalıtsal malzeme): Canlı organizmalarda biyolojik bilgiyi depolayan molekül
- Gen: Organizmada spesifik bir işlevi gerçekleştirmek için gerekli bilgiyi içeren DNA bölgesi
- Genom: Bir organizmanın tüm genlerini içeren paket
- Protein: Gen ifadesi denilen işlem ile gendeki bilgiyi kullanarak üretilen (amino asitlerden oluşan) molekül
- Proteinler bir organizmanın yaşamındaki tüm görevleri gerçekleştirirler
- Genler ifadeleninceye kadar işe yaramazdır
- Gen/protein dizisi: Yapı taşlarının (genler için nükleotidler ve proteinler için amino asitler) ardışık sıralamasından oluşmuş diziler

Gen (nükleik asit) dizisi : “ATCGGTCAGAGTAAC”

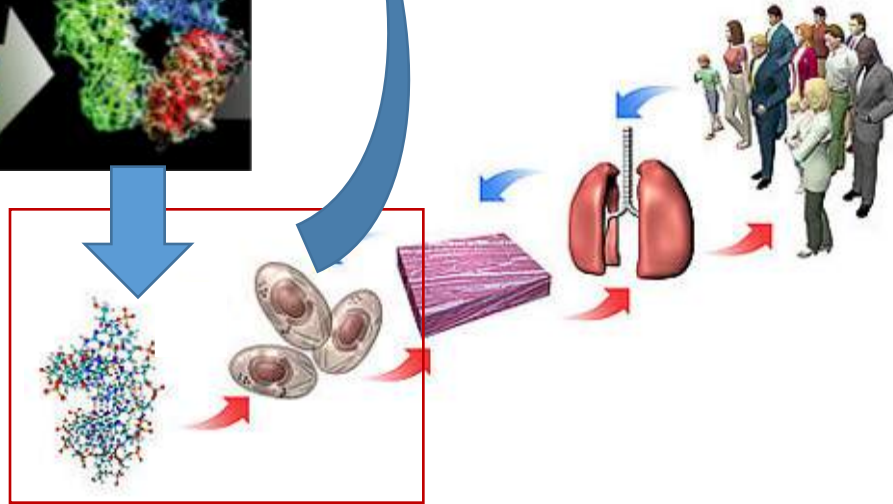
Protein (amino asit) dizisi : “MDFFVRLARE TGDRKREFLE LGRKAGRFP A”

Biyolojik Organizasyon

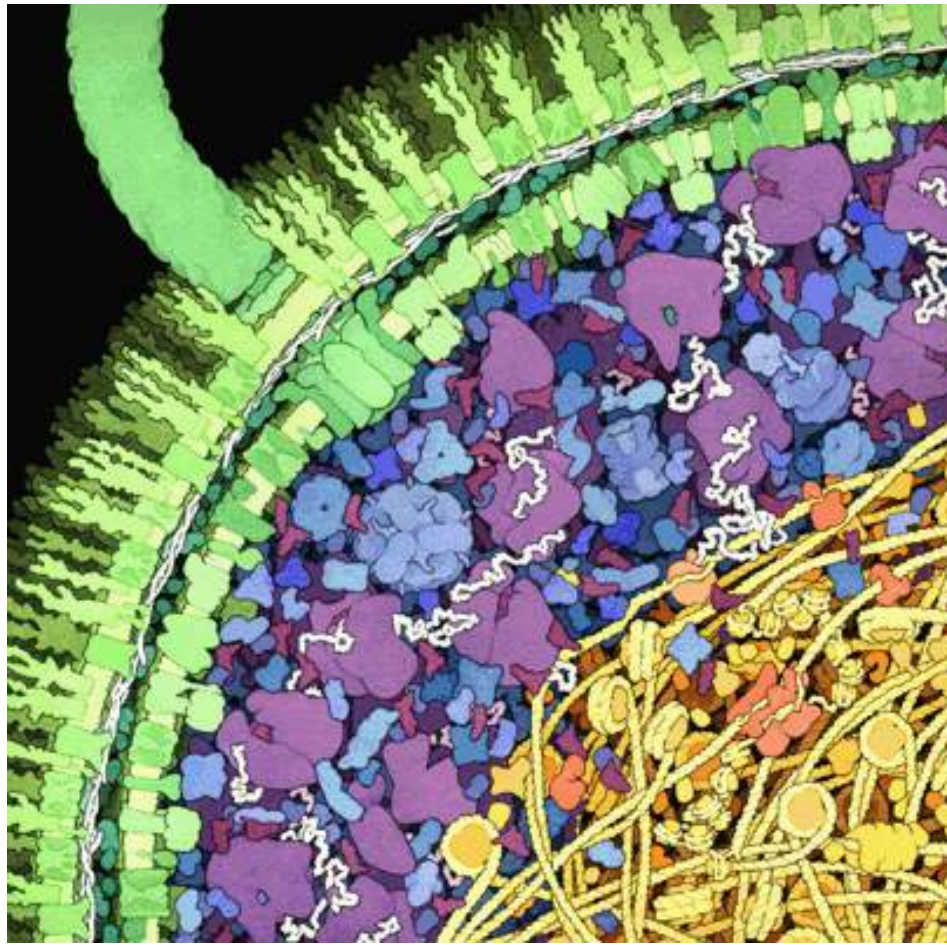


<https://en.wikipedia.org/wiki/Omics>

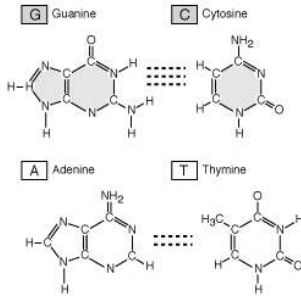
Hücre içerisinde
moleküller arasındaki
değişim ve etkileşim



David
Goodsell,
2000



Bir escherichia coli hücresinin küçük bir kısmı-çok kalabalık



Genlerimizin alfabetinde dört harf mevcut:
A, C, G and T.

Bu harfler nasıl yaşayacağımızı, nasıl büyüyeceğimizi, nasıl
öleceğimizi belirler.



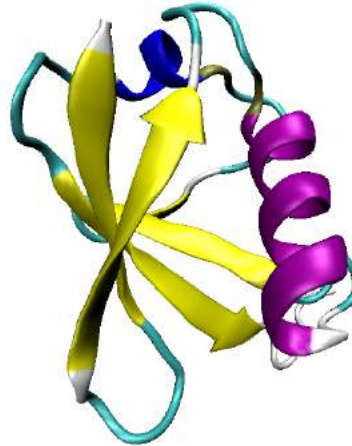
Biyolojik süreçleri anlamak için atomik düzeyde çalışmamız gerekir

TAGTTCCGTCGCAGCCGGGATTTGGGTGCGGTTCTTGTGTGGATCGTGTGATCGTCACTTGACAA TG CAGATCTTCTGTGAA GACTCTGA CTGGTAAGACCATCACCC TC
 GAGGTGAGCCCA GTGACACCATCAGAA TGTCAAAGGCAAAGATC CAAATAAAGGAGCGATCCTCCTGAC CAGCA GAGGCTGATCTTTGCTGGAAAA CAGCTGGAGATGG
 GCGCACCCCTGTCTGACTACAA CATCCAGAAGAGTCCACCTGCACTGGTGC TCCTGTCAGAGGTGGGATGCAATCTTCTGTGAGAGACATCCTACTGGCAA GACCATCACTCT
 TGA GGTGAGC C CAGTGA CACCATCGAGAACGTC AAA GCAAAGATCCAGGACAGGAAGGCATTCCTCTGACCA GCAGAGGTGATCTTTGCTCGGAAA G CAGCTGGAA GATG
 GGC GCACCTGTCGACTACAACATCAGAAA GAGTCTACCTGCA C C TGGTGCTCCTGTCTCAGAGGTGGGATGCAGATCTTCTGTGAAGACCTGACTGTAA GACCATCA C C
 CTCAGGTGGAGCOCAGTGACACCATCAGAA TGTCAAGGCAAAGATC CAA GATAAGGAGCGATTCCTCTCTGATCAGCA GAGGTGATCTTTGCTCGGAAA CAGCTGGAA GA
 TGGTCTGTA C C C TGTCTGACTACAACATCAGAAA GAGTCCACCTTGCACCTGGTACCTCCTGTCTCAGAGGTGGGATGCAATCTTCTGTGAAGACACTCAGTGCAAGACCATCA
 C C C T T GAGT C GAGC C CAGTGACACTATCAGAA C GTCAAAGCAAAGATC CAA GACAA GAGGCA TCTCTCTGAC CAGCAGAGGTGATCTCTTTGCCGGAAGCAGCTGGA A
 GATGGGCGCACCTGTCTGACTACAACATCAGAAA GAGTCTACCTGCA C C TGGTGCTCCTGTCTCAGAGGTGGGATGCA GATCTTCTGTGAAGAC

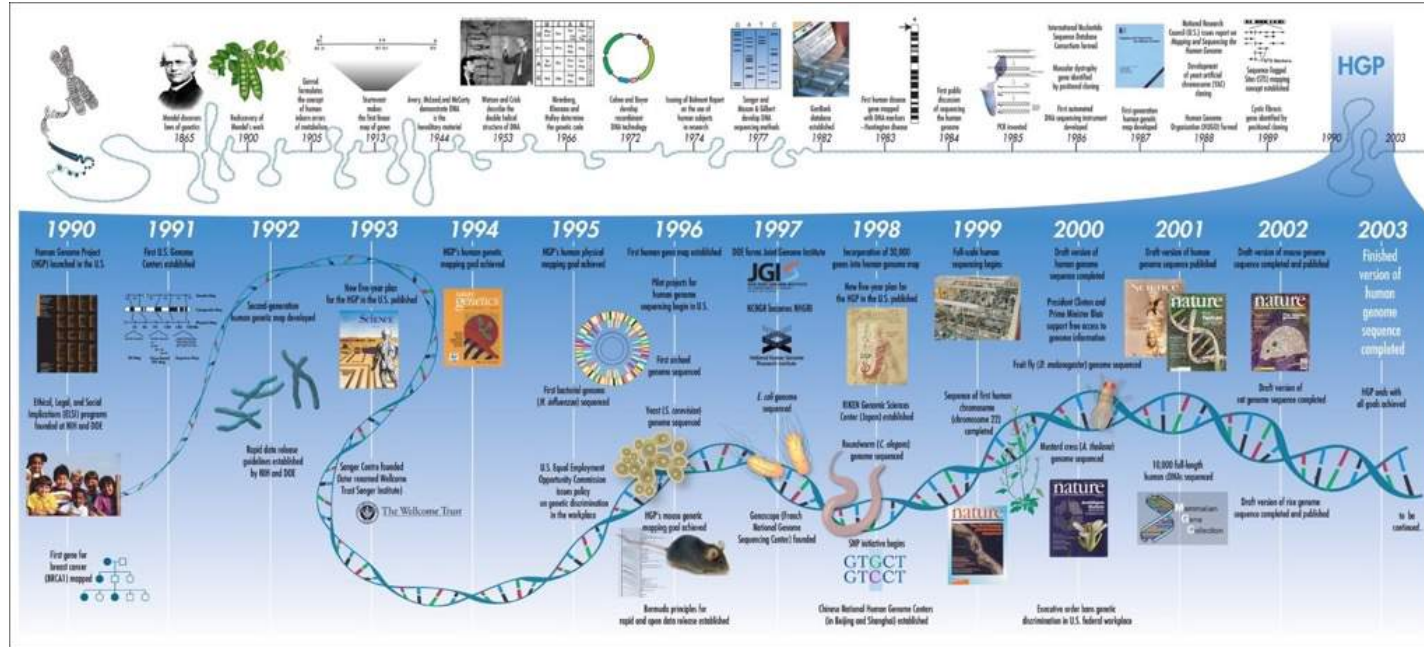


4 harfli alfabe proteinin 20 harfli alfabesine çevrilir

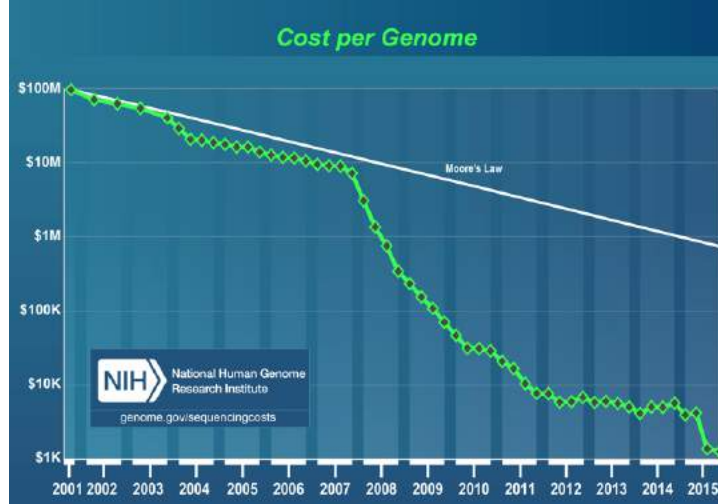
MQIFVKTLTGKTI TLEVEPSDTIENVKAKIQDKEGIPPDQORLIFAGKQLEDGRTLS DYNIQKES TLHLVLRL RGG



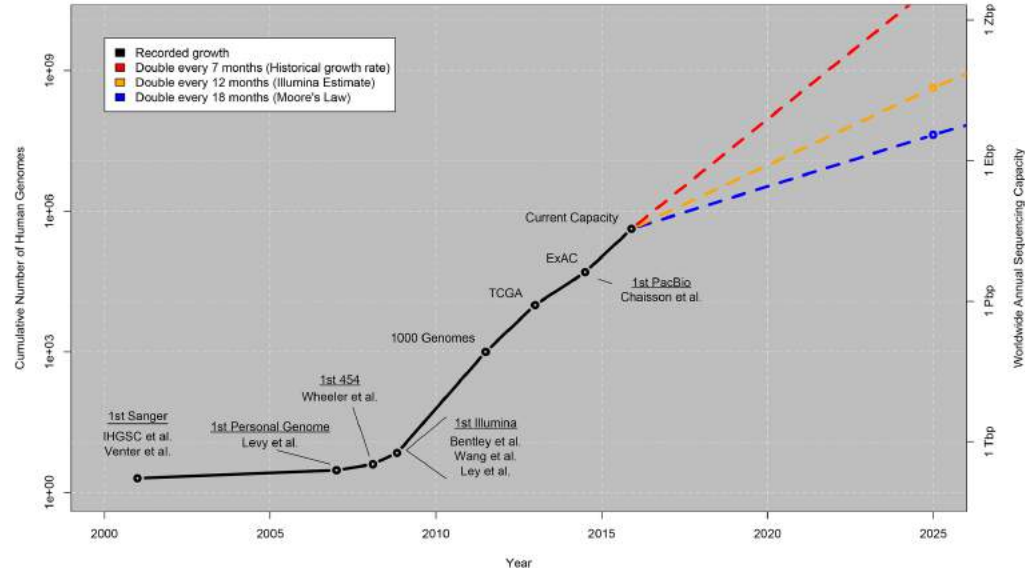
Biyolojik veri birikiminde kilometre taşları ve İnsan Genom Projesi



Dizilenmesi yapılmış verinin birikme hızı

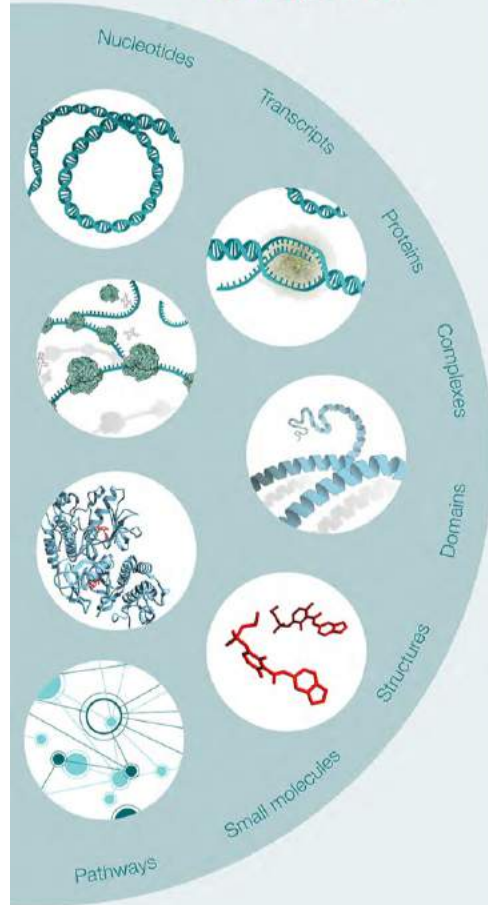


Growth of DNA Sequencing



Stephens Z.D., et al. (2015) Big
Data: Astronomical or Genomical?
PLoS Biol 13(7): e1002195

Data deposition



Integration

Human beings



Tissues & organs

↓
Cells

↓
Biobanks

Translation

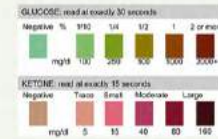
Human populations



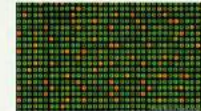
New medicines



Disease prevention



Early diagnosis



Ekibimiz



Resimde olmayanlar

Ö. Sinan Saraç
Ayşe Gül Yaman
Alperen Dalkıran

EBI

Maria Martin
Rabie Saidi

Heval Ataş

Rengül
Çetin-Atalay

Volkan Atalay

Ahmet
Rıfaioğlu

Tunca Doğan

Biyoenformatik ve Dizi Analizi

Giriş

Biyoenformatik: Biyolojik veri analizi için yöntemler, araçlar ve servisler geliştirilmesi



- Gen ve protein dizileri
- Gen ifade verisi
- Protein-protein etkileşimleri
- Histolojik görüntüler

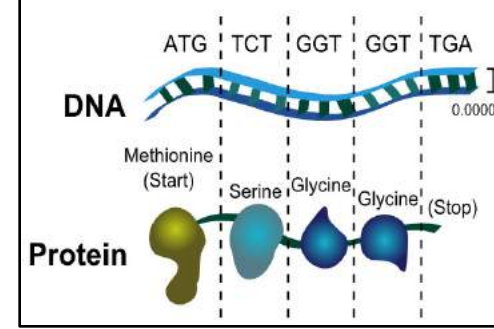


Büyük miktarlarda biriken veriyi analiz etmek için sofistike yöntemler gerekli

İşlemsel biyoloji: Bilimsel çıkarımlar yapabilmek için biyolojik verinin işlemsel analizi

Gen ve protein dizi analizi

- Biyolojik dizilerde birçok sinyal (işaret) gömülüdür
 - Örneğin: işlev, aktivite bölgesi, içinde yer aldığı metabolik süreç
- Bu örüntüler benzer işleve sahip ve geçmiş paylaştan (evrimsel ilintilik taşıyan) proteinlerde ortak olabilir
- Bu tür örüntülerin belirlenmesi yeni bulunan proteinlerin özelliklerinin bulunması açısından önemlidir
- Yapay öğrenme yöntemleri kullanmadan bu örüntüleri bulmak kolay değildir



Source:

<http://www.mynortherndiary.com/diary/science-technology/designing-of-proteins/>

Protein 1: IGPGEDEYCTARNQYGEAICSVYIQPEGAPMPALQPIQNLEKNIYSNG

Protein 2: YSYTSEEEFRVDTFEYRLLREVSFREAITRRSGYEQDSQLSRNQYGEQELDRNQG

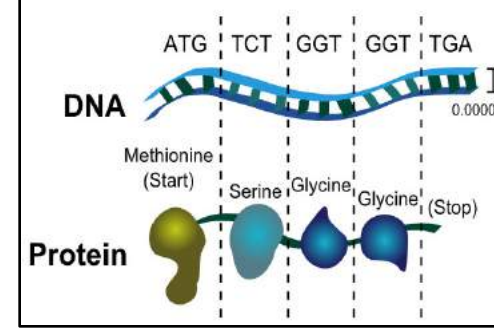
Protein 3: PARNQYGEQAPQISQKPRSSKLIEGSDAVFTARVGS

Protein 4: TLRVKNATARDGGHYTLAENLQGRNQYGE CVSAVLVEPA

Protein 5: AYEPKPVDVMAEQRNQYGELEAGKALPPAFVKAFGDREITEGRMTR

Gen ve protein dizi analizi

- Biyolojik dizilerde birçok sinyal (işaret) gömülüdür
 - Örneğin: işlev, aktivite bölgesi, içinde yer aldığı metabolik süreç
- Bu örüntüler benzer işleve sahip ve geçmiş paylaştan (evrimsel ilintilik taşıyan) proteinlerde ortak olabilir
- Bu tür örüntülerin belirlenmesi yeni bulunan proteinlerin özelliklerinin bulunması açısından önemlidir
- Yapay öğrenme yöntemleri kullanmadan bu örüntüleri bulmak kolay değildir



Source:
<http://www.mynortherndiary.com/diary/science-technology/designing-of-proteins/>

Protein 1: IGPGEDEYCTCTARNQYGEAICSVYIQPEGAPMPALQPIQNLEKNIYSNG

Protein 2: YSYTSIEEEFRVDTFEYRLLEVSFREAITRRSGYEQDSQLSRNQYGEQELDRNQG

Protein 3: PARNQYGEQAPQISQKPRSSKLIEGSDAVFTARVGS

Protein 4: TLRVKNATARDGGHYTLAENLQGRNQYGE CVSAVLVEPA

Protein 5: AYEKPVDVMAEQRNQYGELEAGKALPPAFVKAFGDREITEGRMTR

Gen ve protein dizi analizi ve anlamlandırması

- Amaç: Henüz anlaşılmamış proteinlerin özelliklerinin benzerlerine bakarak belirlenmesi
- Veri: nükleik/amino asit dizileri, fizikokimyasal özellikler, işlevsel anlamlandırmalar
- Yöntemler: yapay öğrenme, sınıflandırma, kümeleme, öznitelik seçimi, boyut indirgeme
- Dizi hizalama: iki veya daha fazla dizinin paylaştığı bölgelerin bulunması-en temel dizi analizi yöntemi
- Popüler hizalama araçları: BLAST (ikili hizalama), Clustal series (çoklu dizi hizalama)

GAATTCAG	GAATTCAG
GGA-TC-G	GCAT-C-G
GAATTC-A	GAATTC-A
GGA-TCGA	GCAT-CGA

Kaynak:

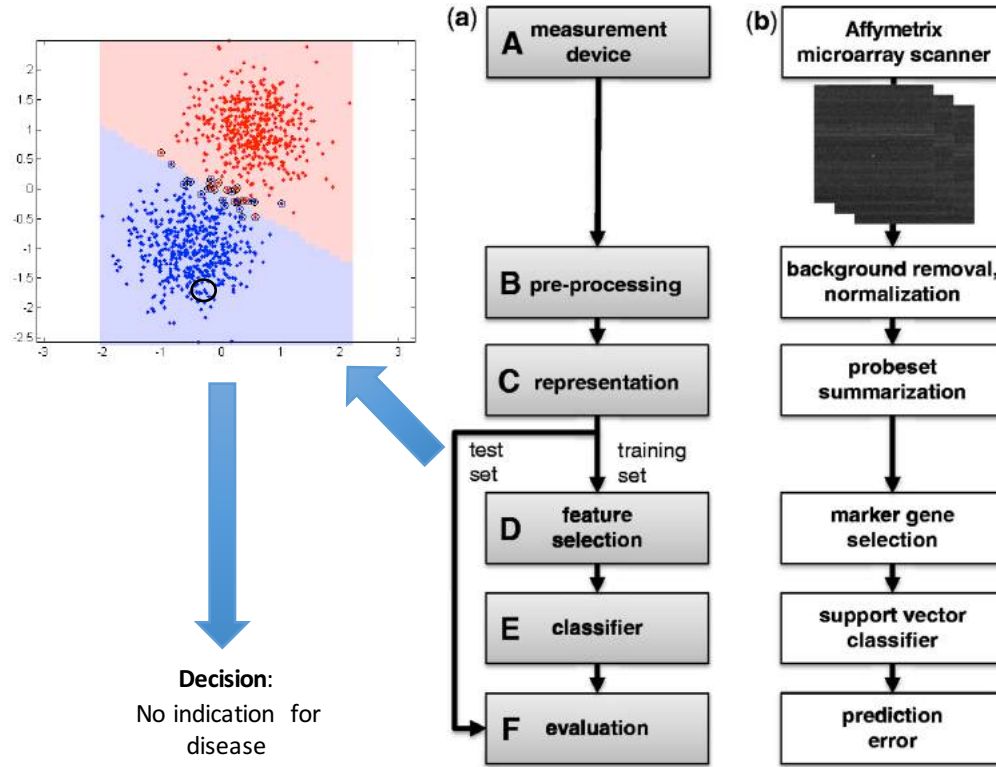
http://statweb.stanford.edu/~nzhang/345_web/

```
A5ASC3.1 14 SIKLWPPSQTLLVERMANNLST..PSIFTRK.
B4F917.1 13 SIKLWPPSESTRIMLVDRMTNNLST..ESIFSRK.
A9S1V2.1 23 VKLWPPSQGTREAVRDKMALKLS..ACFESQS.
B9GSN7.1 13 VKLWPPSQSTRMLMVERMTKNFIT..PSFISRK.
Q8H056.1 30 SFSIWPPQTQTRDAVVRRLVDTLGG..DTILCKR.
Q0D4Z3.2 44 SLSIWPPSQTRDAVVRRLVQTLVA..PSILSKR.
B9HVV8.1 56 SFSIWPPQTQTRDAITISRLIETLST..TSVLSKR.
Q0IYC5.1 29 SFAWPPTRRTQDAVVRRLVAVLGGDTTALRKRY.
A9NW46.1 13 SIKLWPPSESTRMLMVERMTDNLSS..VSFFSRK.
Q9C500.1 57 SLRIWPPQTQTRDAVLRRLIETLST..ESILSKR.
Q2HRI7.1 25 NYSIWPPKQTRDAVKNRLIETLST..PSVLTQR.
Q9M7N3.1 28 SFKIWPPQTQTRDAVVRRLVETLTS..QSVLSKR.
Q9M7N6.1 25 SFSIWPPQTQTRDAVINRLIESLST..PSILSKR.
Q9LE82.1 14 SVKMWPPSKSTRMLMVERMTKNIT..PSIFSRK.
Q9M651.2 13 SIKLWPPSLPTRKALIERITNNFSS..KITIFTEK.
B9R748.1 48 SLSIWPPQTQTRDAVITRLIETLSS..PSVLSKR.
```

Kaynak:

https://en.wikipedia.org/wiki/Sequence_analysis#/media/File:WPP_domain_alignment.PN
G

(A) Sınıflandırma için izlenmesi gereken işlemler.



Dick de Ridder et al. Brief Bioinform 2013;14:633-647

İlgi duyanlar için yararlı yayınlar:

- Baldi, P., & Brunak, S. (2001). *Bioinformatics: the machine learning approach*. MIT press.
- Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., ... & Robles, V. (2006). Machine learning in bioinformatics. *Briefings in bioinformatics*, 7(1), 86-112.
- de Ridder, D., de Ridder, J., & Reinders, M. J. (2013). Pattern recognition in bioinformatics. *Briefings in bioinformatics*, 14(5), 633-647.
- Friedberg, I. (2006). Automated protein function prediction—the genomic challenge. *Briefings in bioinformatics*, 7(3), 225-242.

Diziden Öngörü

Anlamlandırma

- Biyolojik dizilerin anlamlandırılması (sequence annotation)
DNA, RNA veya protein dizilerinin özgül özelliklerinin yapı veya işlev hakkında betimleyici bilgi ile işaretlenmesi işlemi
 - Proteinlerin işlevlerinin bilinmesi
 - kanser gibi ölümcül süreçlerin önlenmesi ya da durdurulmasından,
 - her türlü hastalık için ilaç tasarımı
- konularına kadar çok geniş bir yelpazede vazgeçilmez öneme sahiptir.

Anlamlandırma

- Protein dizi anlamlandırması (protein sequence annotation)
- Protein dizisinden işlev anlamlandırması (function annotation)
- Protein dizisinden işlev öngörülmesi (function prediction)

Sunum İçeriği

- Çok sayıda proteinin işlevlerinin in silico (bilgisayar kullanarak) öngörmek için
 - geliştirdiğimiz yapay öğrenme ve derin öğrenme **yöntemleri**,
 - eğitim ve sınav **veri kümelerinin** oluşturulması,
 - **başarımlarının** değerlendirilmesi ve standart veri kümeleri üzerinde **karşılaştırmalarının** yapılması,
 - **servis** olarak web üzerinden sunulması veya indirilebilir kodlar ve veri.

Proteinler

- Proteinler büyük moleküllerdir.
- Proteinler amino asitlerden oluşur.
- Proteinler birçok işlevi yerine getirir; örneğin:
 - biyokimyasal reaksiyonların katalize edilmesi
 - DNA'nın kopyalanması
 - hürelerearası taşıma
 - vücudun virüs ve bakterilerden korunması
- Binlerce işlev var-hangileri?



2PYQ proteinin yapısı
<http://godziklab.org/SSBC/modeling.html>

Protein işlevi ve yapısı

- Proteinin işlevini belirleyen yapısıdır
- Yapısı belirlenmiş protein sayısı: 141.616 (PDB 2.Temmuz.2018)
- Dizisi bilinen protein sayısı: 116.030.110 (TrEMBL 2.Temmuz.2018)
- Diziden işlevi öngör

Problem

Problem tanımı

- Sadece az sayıda proteinin moleküler özellikleri deneyler sonucunda belirlenebilmiş (pahalı, emek yoğun ve zaman alan bir iş)
- Gün be gün daha fazla genom dizileniyor, sunucularda veri birikiyor
- Bu bilgilerin kullanılabilmesi için yeni dizilerin özellikleri belirlenmesi gerekiyor (örneğin, hastalıklara tedavi bulmak veya yeni biyoteknolojik ürün geliştirmek için)
- “protein işlev öngörüsü” problemi

Veri:

- Proteinlerin amino asit dizileri
- Bazı proteinlerin işlevsel anlamlandırmaları

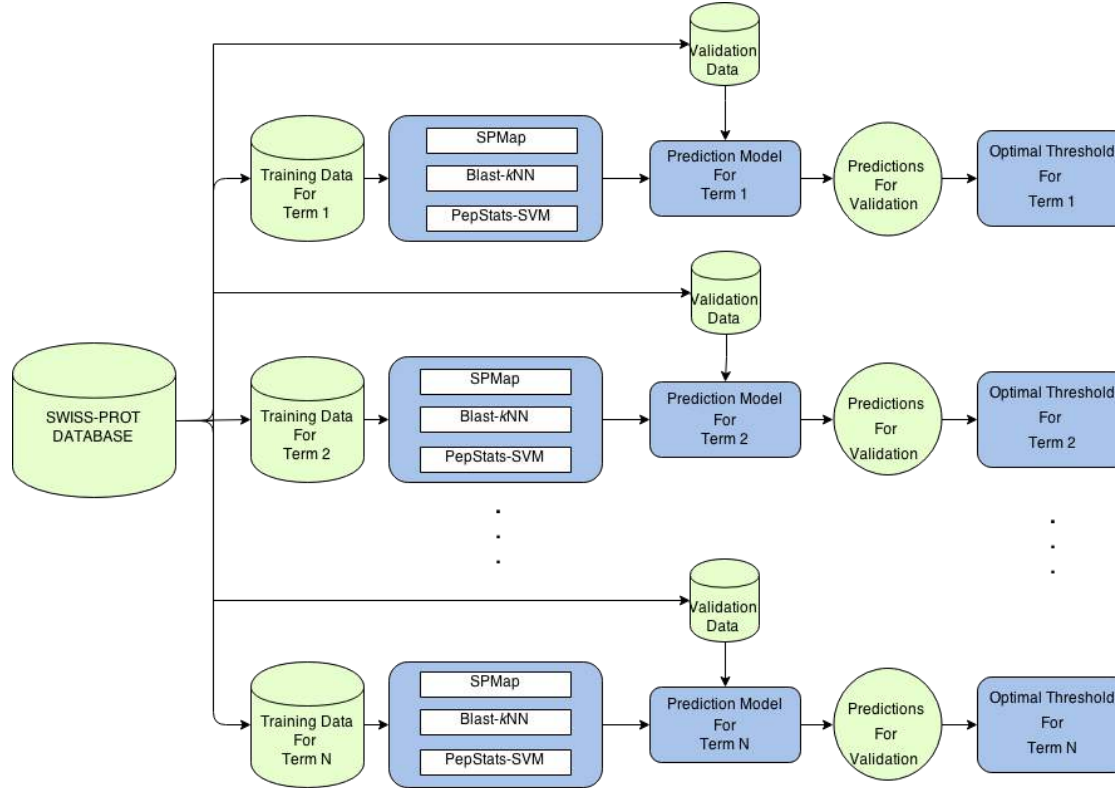
Yöntem yaklaşımı:

- Yapay öğrenme-gözetimli yaklaşım (sınıflandırma)
- Eğitim için kullanılacak kıt sayıda güvenilir etiketli veri
- Özellikleri bilinmeyen dizileri sorgula

Diziden İşlev Öngörmek-Zorluklar

- Doğrudan öngörülemiyor
- Farklı uzunluklar
- Vektör gösterimi bulmak-sabit uzunluk
- Çok fazla işlev var

Eğitme sistemi-genel bakış



Başlıklar-tekrar

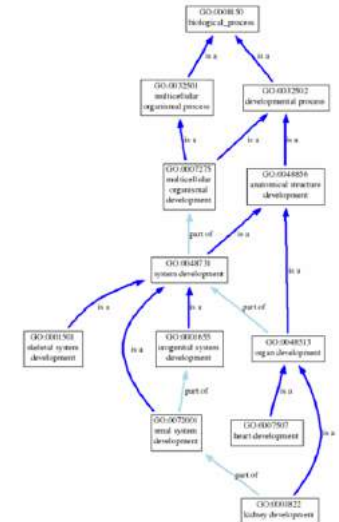
1. Sınıflandırıcı oluşturmak
2. Eğitim, doğrulama ve sınama veri kümeleri oluşturmak
3. Başarım, karşılaştırma
4. Servis sunumu

Protein işlevinin sistematik tanımı

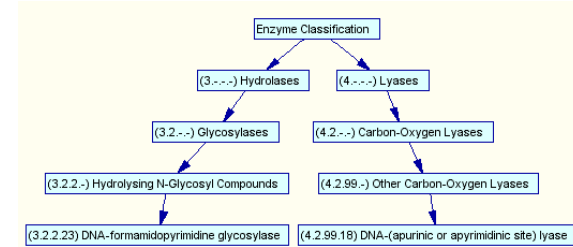
- Soyut bir kavram olmasından dolayı protein işlevinin işlemsel bir sisteme dahil edilmesi kolay değil

Protein	Cellular tumor antigen p53
Gene	TP53
Organism	Homo sapiens (Human)
Status	Reviewed - Annotation score: ★★★★★ - Experimental evidence at protein level
Function	Acts as a tumor suppressor in many tumor types; induces growth arrest or apoptosis depending on the physiological circumstances and cell type. Involved in cell cycle regulation as a trans-activator that acts to negatively regulate cell division by controlling a set of genes required for this process. One of the activated genes is an inhibitor of cyclin-dependent kinases. Apoptosis induction seems to be mediated either by stimulation of BAX and FAS antigen expression, or by repression of Bcl-2 expression. In cooperation with mitochondrial PPF is involved in activating oxidative stress-induced necrosis; the function is largely independent of transcription. Induces the transcription of long intergenic non-coding RNA p21 (lincRNA-p21) and lincRNA-Mkin1. LincRNA-p21 participates in TP53-dependent transcriptional repression leading to apoptosis and seem to have to effect on cell-cycle regulation. Implicated in Notch signaling cross-over. Prevents CDK7 kinase activity when associated to CAK complex in response to DNA damage, thus stopping cell cycle progression. Isoform 2 enhances the transactivation activity of isoform 1 from some but not all TP53-inducible promoters. Isoform 4 suppresses transactivation activity and impairs growth suppression mediated by isoform 1. Isoform 7 inhibits isoform 1-mediated apoptosis. Regulates the circadian clock by repressing CLOCK-ARNTL/BMAL1-mediated transcriptional activation of PER2 (PubMed:24051492). 12 Publications

- Bu soruna çözüm için ve proteinlerin niteliklerini tanımlamak amacıyla kontrollü kelime hazneleri (ontolojiler) önerilmiş
- Gen Ontolojisi (GO)** en popüler olanı: işlevlerin genelden öze doğru gösterildiği yönlendirilmiş asıkk çizge (DAG)
- Enzim Komisyonu (EC)** enzimlerin katalize ettikleri reaksiyolara göre işlevsel organizasyonel sistem



Kaynak: <http://wacławikgen677s10.weebly.com/gene-ontology.html>



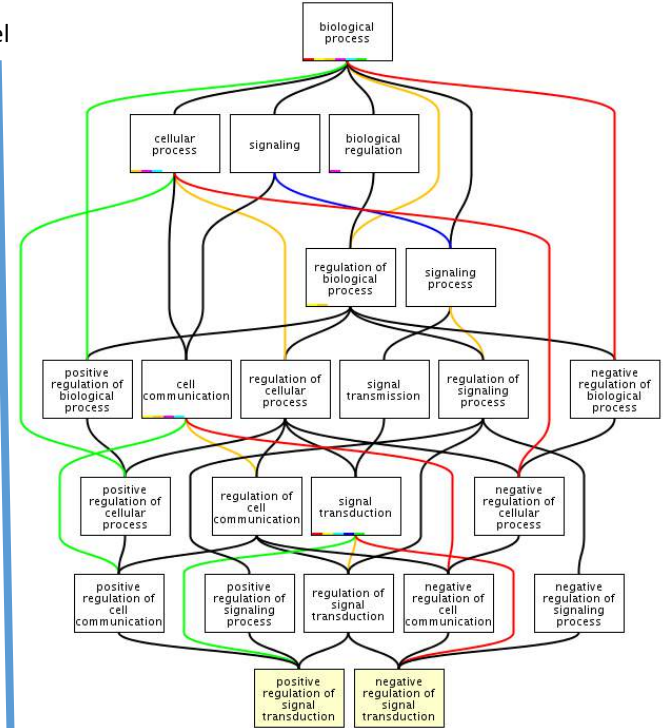
Kaynak: <http://www.enzyme.chem.msu.ru/hcs/distributions/P-EC-directly-linked-proteins.html>

Proteinin İşlevleri

GEN ONTOLOJİSİ (GO) <http://www.geneontology.org/>

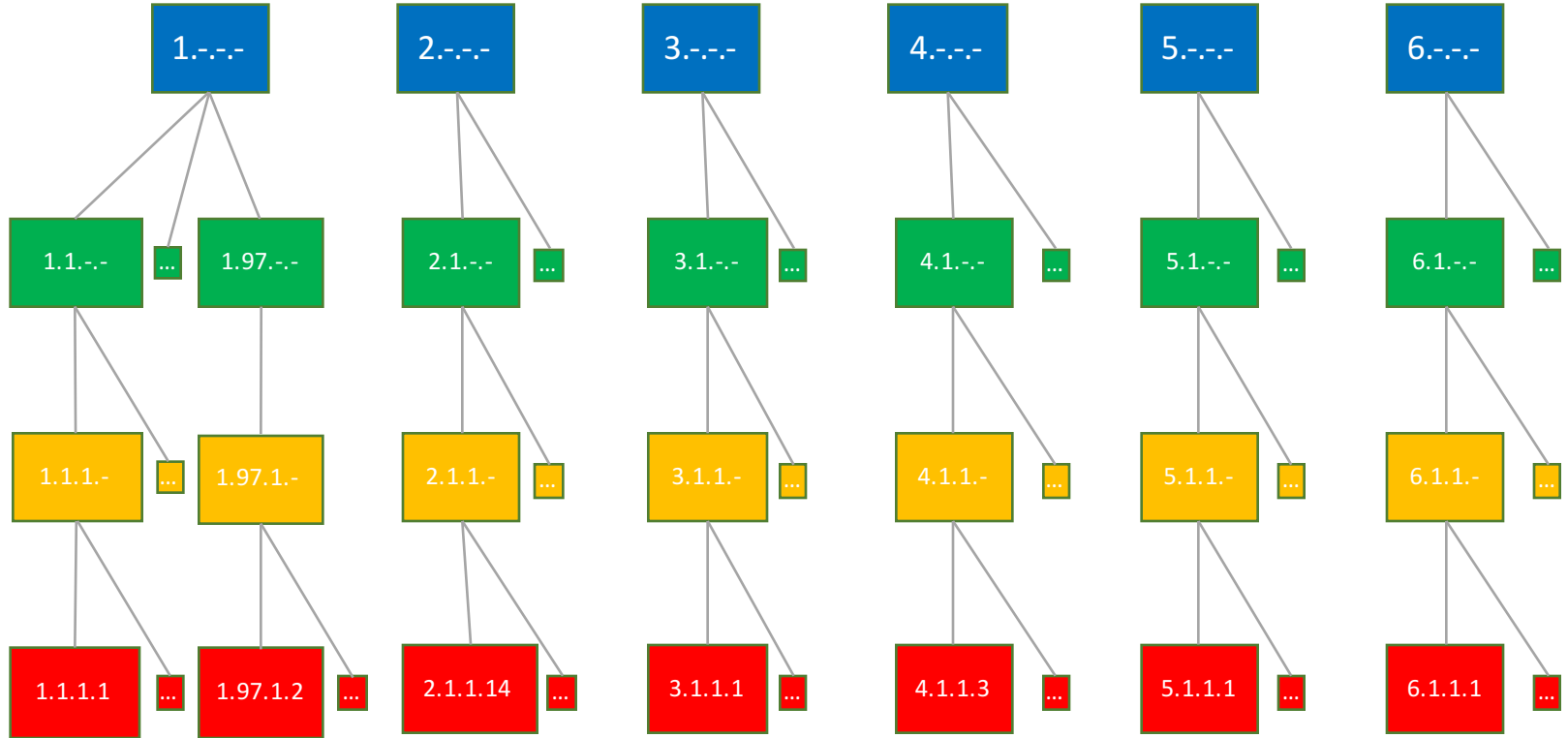
- Kontrollü kelime haznesi (controlled vocabulary)
- Alan bilgisi
- İlişkilerle birbirine bağlanmış GO terimleri
- Üç kategori
 - Moleküler işlev (molecular function): bir genin işlevleri
 - Biyolojik süreç (biological process): birden fazla gen ürününün aktivitelerinden oluşan daha büyük süreçler ve yollar
 - Hüresel bileşen (cellular component): gen ürünlerinin aktif oldukları yer

Genel



Özel

EC Ağacı



Başlıklar-tekrar

1. Sınıflandırıcı oluşturmak
2. Eğitim, doğrulama ve sınama veri kümeleri oluşturmak
3. Başarım, karşılaştırma
4. Servis sunumu

Sınıflandırıcı

- Binlerce GO terimi: tek bir sınıflandırıcı ile hepsini öngörmek zor
 - Herbir terim için ikili sınıflandırıcı
 - 1 işlev var
 - 0 işlev yok
 - Birbirinden bağımsız sınıflandırıcılar
 - Çoklu sınıflandırıcı (multi-task)-birden fazla terim
 - Daha sonra

Eğitim ve sinama veri kümeleri

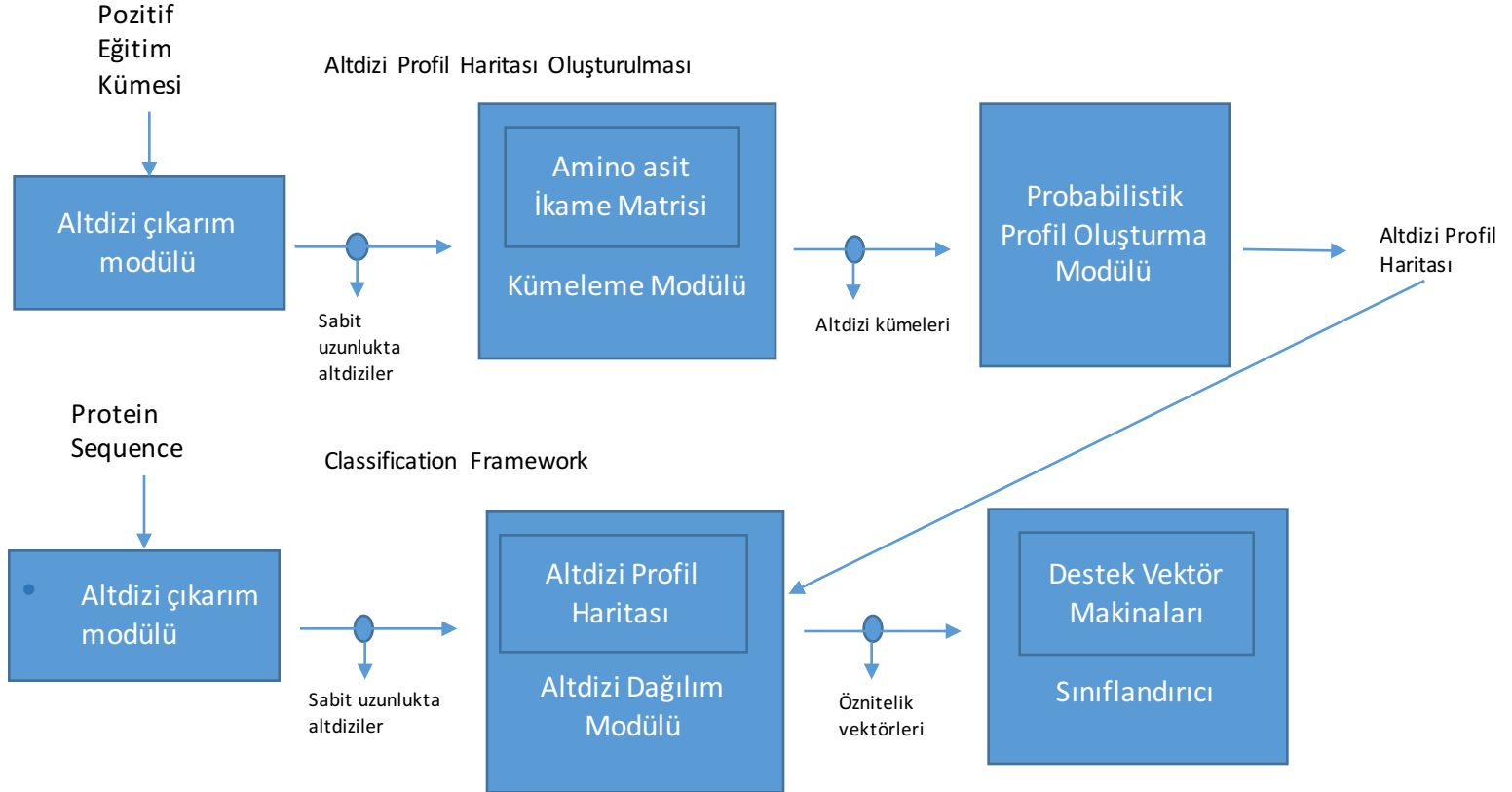
- Veri kümesi
 - Örnek: %80 eğitim ve geçerleme, %20 sinama
 - %80 eğitim ve geçerleme -> 4 veya 5 katmanlı çapraz geçerleme
 - Eğitim veri kümesi: öğrenilebilen parametrelerin belirlenmesi
 - Geçerleme veri kümesi: öğrenilemeyen parametrelerin seçimi
 - %20 sinama
 - Başarım ölçümü
 - Ayrıca tamamen bağımsız sinama veri kümeleri
 - Hangi model son olarak kullanılacak?

Eğitim Veri Kümeleri-ikili sınıflandırıcı

- Pozitif veri kümesi
 - İşleve sahip olarak etiketlenmiş protein dizileri
- Negatif veri kümesi
 - İşleve sahip olmayan protein dizileri
 - Veritabanlarında açıkça belirtilmiyor, raporlanmıyor
 - Kolay yolu: pozitif veri kümesi dışında kalan herşey
 - Sayı çok
 - Şimdiye kadar etiketlenmemiş olması, pozitif etiketlenmeyeceği anlamına gelmiyor
 - En uzak olasılıkları seç

Sınılandırıcı Oluřturma

Sınıflandırıcı-SPMAP



Diger Teknik Detaylar

- Vektör gösterimi
- Kodlama (encoding)

Vektör gösterimi

- Sorun: dizi uzunlukları 50-5000 arasında değişiyor
- Seçenekler
 - Amino asit kompozisyonu: herbir amino asitin sıklığı-20 boyutlu vektör
 - Sorun: dizi içindeki sıra, sıralama bilgileri kayboluyor
 - Amino asit ikilileri: $20 \times 20 = 400$ boyutlu vektör
 - k sayıda yanyana gelmiş amino asit altdizileri: 20^k boyutlu vektör
 - Çoğunluğu 0, boş

Örnek k -mer

$k=5$

MSTNPKPQR

MSTNP

STNPK

TNPKP

NPKPQ

PKPQR

- k -mer = altdizi (subsequence)

k -mer'ler

- k tane amino asit içeren altdiziler: k -mer'ler (k -mers)
- Uzay: eğitim veri kümesinin k -mer'leri (tüm olası k -mer'ler yerine)
 - Bu uzayda k -mer'lerin dağılımı
- Uzayı daraltalım
 - eğitim veri kümesinin k -mer'lerini kümeleyelim (clustering)
 - Aynı kümeye düşen ve birbirine benzeyen k -mer'leri bir prototip k -mer ile gösterebiliriz

k-mer'ler ve benzerlik

- Pozitif eğitim veri kümesinde olan protein dizilerinden *k*-mer'ler çıkart
- Çıkartılmış olan *k*-mer'leri kümele
- Kümelemek için iki *k*-mer, *x* ve *y* arasında benzerlik tanımlayalım

$$s(x, y) = \sum_{k=1}^5 M(x(k), y(k))$$

$x(k)$: *x* *k*-mer'inin *k*. pozisyonundaki amino asit

Örneğin, *x* = MSTNP and *y* = STNPK

$$s(x, y) = M(M, S) + M(S, T) + M(T, N) + M(N, P) + M(P, K)$$

İki amino asitin benzerliği

BLOSSUM62

- İkame matrisi
- Dizilerin hizalanması için kullanıyor
- Herbir eleman: iki amino asit arasındaki benzerlik skoru

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1
F	-2	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1

k-mer'ler ve benzerlik

$$s(x, y) = \sum_{k=1}^5 M(x(k), y(k))$$

$x(k)$: x k -mer'inin k . pozisyonundaki amino asit

Örneğin, $x = \text{MSTNP}$ and $y = \text{STNPK}$,

$$\begin{aligned} s(x, y) &= M(M, S) + M(S, T) + M(T, N) + M(N, P) + M(P, K) \\ &= (-1) + 1 + 0 + (-2) + -1 \\ &= -3 \end{aligned}$$

BLOSSUM62 kullanınca

Kümeler ve benzerlik

c kümesi ve s altdizisi arasındaki benzerlik hesaplandıktan sonra,

- Eğer $s(c,s) \geq 8$, s altdizisi c kümesine atanır
- Eğer $s(c,s) < 8$, yeni bir küme oluşturulur

PSSM

- Herbir küme için
 - Herbir pozisyondaki amini asitlerin sıklığını bul (PSSM)
 - Ve dolayısıyla bu pozisyonda bulunma olasılıklarını çıkart (probability matrix)
- Herbir küme için pozisyona özel skor matrisi (position specific scoring matrix-PSSM) oluşturulur
 - 5 kolon (k -mer sayısı) ve 20 sıra (amino asit sayısı)

	1	2	3	4	5
A	0	0	0	0	0
R	0	0	0	0	0
N	0	0	0	1	0
D	0	0	0	0	0
C	0	0	0	0	0
Q	0	0	0	0	0
E	0	0	0	0	0
G	0	0	0	0	0
H	0	0	0	0	0
I	0	0	0	0	0
L	0	0	0	0	0
K	0	0	0	0	0
M	1	0	0	0	0
F	0	0	0	0	0
P	0	0	0	0	1
S	0	1	0	0	0
T	0	0	1	0	0
W	0	0	0	0	0
Y	0	0	0	0	0
V	0	0	0	0	0
	M	S	T	N	P

A	11	1	20	2	0
R	2	2	1	0	108
N	0	0	0	0	0
D	0	0	0	0	0
C	0	0	0	0	0
Q	0	0	0	0	1
E	1	3	0	2	0
G	0	94	0	0	2
H	0	0	0	0	1
I	4	0	76	62	2
L	5	0	13	8	2
K	2	4	0	0	6
M	100	1	1	1	0
F	1	0	0	0	0
P	2	84	0	1	1
S	16	1	0	2	0
T	30	0	0	93	59
W	0	0	0	0	0
Y	5	0	0	0	0
V	11	0	79	19	8

$S_i = 190$



A	-2.84	-5.23	-2.25	-4.54	-9.85
R	-4.54	-4.54	-5.23	-9.85	-0.56
N	-9.85	-9.85	-9.85	-9.85	-9.85
D	-9.85	-9.85	-9.85	-9.85	-9.85
C	-9.85	-9.85	-9.85	-9.85	-9.85
Q	-9.85	-9.85	-9.85	-9.85	-5.23
E	-5.23	-4.14	-9.85	-4.54	-9.85
G	-9.85	-0.70	-9.85	-9.85	-4.54
H	-9.85	-9.85	-9.85	-9.85	-5.23
I	-3.85	-9.85	-0.91	-1.12	-4.54
L	-3.63	-9.85	-2.58	-3.16	-4.54
K	-4.54	-3.85	-9.85	-9.85	-3.45
M	-0.64	-5.23	-5.23	-5.23	-9.85
F	-5.23	-9.85	-9.85	-9.85	-9.85
P	-4.54	-0.81	-9.85	-5.23	-5.23
S	-2.47	-5.23	-9.85	-4.54	-9.85
T	-1.84	-9.85	-9.85	-0.72	-1.17
W	-9.85	-9.85	-9.85	-9.85	-9.85
Y	-3.63	-9.85	-9.85	-9.85	-9.85
V	-2.84	-9.85	-0.87	-2.30	-3.16

PSSM to Profile

Olasılıksal profil

PSSM ile amino asit sayıları elde edildikten sonra, herbir PSSM olasılıksal profile dönüştürülür.

- $P_k(i, j)$ i . pozisyonda j amino asitinin olma olasılığı
- $Aa_{count}(i, j)$ i . pozisyonda j amino asitinin sayısı
- S_i normalize etmek amaçlı

$$PP_k(i, j) = \log \frac{Aa_{count}(i, j) + 0.01}{S_i}$$

Öznitelik vektörü

- Bir altdizinin herbir profil (küme) tarafından üretilmiş olma olasılığı
- Profil için en yüksek olasılık -> öznitelik vektörünün profile karşılık gelen elemanı
- s altdizisi için

$$P(s|PP_k) = \sum_{i=1}^5 PP_k(i, s(i))$$

Öznitelik vektörü

Öznitelik vektörünün j . elemanı

$$V(j) = \max_{s_i \in E} P(s_i | PP_k)$$

e proteini için PP_k 'da bulunan en yüksek olasılık değeri (altdizi- k -mer)

Öznitelik vektörü

- Vektörün boyutu = küme (cluster) sayısı
- k -mer'leri bul
 - Herbir k -mer için
 - Küme olasılık matrisi ile karşılaştırarak k -mer'in bu kümede bulunma olasılığını bul
 - Tüm kümeler için en yüksek olasılık değeri vektörün o pozisyona karşılık gelen değeri

For example, first profile values for each position:

A	-2.84	-5.23	-2.25	-4.54	-9.85
R	-4.54	-4.54	-5.23	-9.85	-0.56
N	-9.85	-9.85	-9.85	-9.85	-9.85
D	-9.85	-9.85	-9.85	-9.85	-9.85
C	-9.85	-9.85	-9.85	-9.85	-9.85
Q	-9.85	-9.85	-9.85	-9.85	-5.23
E	-5.23	-4.14	-9.85	-4.54	-9.85
G	-9.85	-0.70	-9.85	-9.85	-4.54
H	-9.85	-9.85	-9.85	-9.85	-5.23
I	-3.85	-9.85	-0.91	-1.12	-4.54
L	-3.63	-9.85	-2.68	-3.16	-4.54
K	-4.54	-3.85	-9.85	-9.85	-3.45
M	-0.64	-5.23	-5.23	-5.23	-9.85
F	-5.23	-9.85	-9.85	-9.85	-9.85
P	-4.54	-0.81	-9.85	-5.23	-5.23
S	-2.47	-5.23	-9.85	-4.54	-9.85
T	-1.84	-9.85	-9.85	-0.71	-1.17
W	-9.85	-9.85	-9.85	-9.85	-9.85
Y	-3.63	-9.85	-9.85	-9.85	-9.85
V	-2.84	-9.85	-0.87	-2.30	-3.16

Now we calculate vector for MSTNP. It is $-0.64 + -5.23 + -9.85 + -9.85 + -5.23 = -30.8$

Second profile values for each position:

A	-9.86	-3.17	-4.15	-3.86	-4.56
R	-5.24	-4.15	-5.24	-1.62	-5.24
N	-9.86	-9.86	-9.86	-4.56	-9.86
D	-4.15	-9.86	-9.86	-9.86	-9.86
C	-9.86	-9.86	-4.56	-9.86	-9.86
Q	-9.86	-9.86	-4.15	-3.86	-9.86
E	-2.31	-9.86	-3.17	-2.61	-9.86
G	-9.86	-3.86	-4.15	-5.24	-9.86
H	-9.86	-9.86	-3.86	-5.24	-9.86
I	-5.24	-3.06	-5.24	-9.86	-3.31
L	-9.86	-1.82	-4.56	-3.86	-5.24
K	-9.86	-9.86	-5.24	-2.95	-9.86
M	-5.24	-5.24	-4.56	-0.55	-4.15
F	-9.86	-9.86	-9.86	-9.86	-5.24
P	-0.26	-5.24	-0.53	-5.24	-9.86
S	-9.86	-9.86	-4.56	-4.56	-5.24
T	-2.31	-9.86	-2.21	-9.86	-4.15
W	-9.86	-9.86	-9.86	-9.86	-9.86
Y	-9.86	-9.86	-9.86	-9.86	-4.56
V	-5.24	-0.35	-1.96	-5.24	-0.11

Now we calculate vector for MSTNP. It is $-5.24 + -9.86 + -2.21 + -4.56 + -9.86 = -31.73$

Öznitelik vektörü

- Herbir profil için bu işlem tekrarlanır ve bir öznitelik vektörü elde edilir.

	1	2	3	4	...	...	283	284
MSTNP	-30.80	-31.73	-21.06	-11.15	...	...	-25.22	-10.11
STNPK	-20.12	-15.21	-11.46	-16.23	...	...	-15.26	-19.65
TNPKP	-15.41	-26.96	-19.29	-28.41	...	...	-19.45	-20.99
NPKPQ	-24.16	-24.63	-28.43	-21.56	...	...	-11.09	-16.36
PKPQR	-10.55	-19.58	-30.78	-19.89	...	...	-22.87	-31.21
...								
...								

Now, our vector becomes {-10.55,-15.21,-11.46,-11.15, ..., -11.09,-10.11}

Sınıflandırıcı

- Destek vektör makinası

Diğer sınıflandırıcılar

- BLAST- k NN
- PEPSTATS-SVM

Diğer sınıflandırıcılar: Blast-kNN

- Homolojiye (benzerlik) dayalı
- Hedef protein ile eğitim veri kümesindeki proteinlerin benzerlikleri

$$O_B = \frac{S_p - S_n}{S_p + S_n}$$

S_p pozitif eğitim veri kümesindek en yakın k komşunun skorlarının toplamı

S_n negatif eğitim veri kümesindek en yakın k komşunun skorlarının toplamı

Diğer sınıflandırıcılar: Pepstats-SVM

```

PEPSTATS of MURI_LISMH from 1 to 266

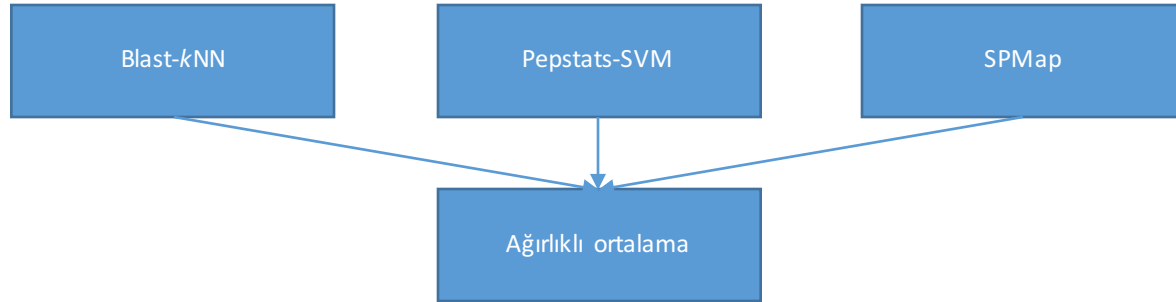
Molecular weight = 29175.76      Residues = 266
Average Residue Weight = 109.683  Charge = -1.5
Isoelectric Point = 6.0474

A280 Molar Extinction Coefficients = 23920 (reduced) 23170 (cystine bridges)
A280 Extinction Coefficients 1mg/ml = 0.786 (reduced) 0.794 (cystine bridges)
Improbability of expression in inclusion bodies = 0.518

Residue   Number   Mole%   DayhoffStat
A = Ala   21       7.895   0.918
C = Cys    4       1.504   0.519
D = Asp   15       5.639   1.025
E = Glu   19       7.143   1.190
F = Phe    8       3.008   0.835
G = Gly   20       7.519   0.895
H = His    5       1.880   0.940
I = Ile   19       7.143   1.587
K = Lys   20       7.519   1.139
L = Leu   26       9.774   1.321
M = Met    7       2.632   1.548
N = Asn    9       3.383   0.787
P = Pro   11       4.135   0.795
Q = Gln    5       1.880   0.482
R = Arg   10       3.759   0.767
S = Ser   14       5.263   0.752
T = Thr   19       7.143   1.171
V = Val   24       9.023   1.367
W = Trp    2       0.752   0.578
Y = Tyr    8       3.008   0.885

Property   Residues   Number   Mole%
tiny       (A+C+G+S+T)  78       29.323
Small      (A+B+C+D+G+N+P+S+T+V) 137       51.504
Aliphatic   (A+I+L+V)    90       33.835
Aromatic    (F+H+W+Y)    23       8.647
Non-polar   (A+C+F+G+I+L+M+P+V+W+Y) 150       56.391
Polar       (D+E+H+K+N+Q+R+S+T+Z) 116       43.609
Charged     (B+D+E+H+K+R+Z)  69       25.940
Basic       (H+K+R)      35       13.158
Acidic      (B+D+E+Z)    34       12.782
    
```

Sınıflandırıcıların birleştirilmesi



Ağırlıkların öğrenilmesi

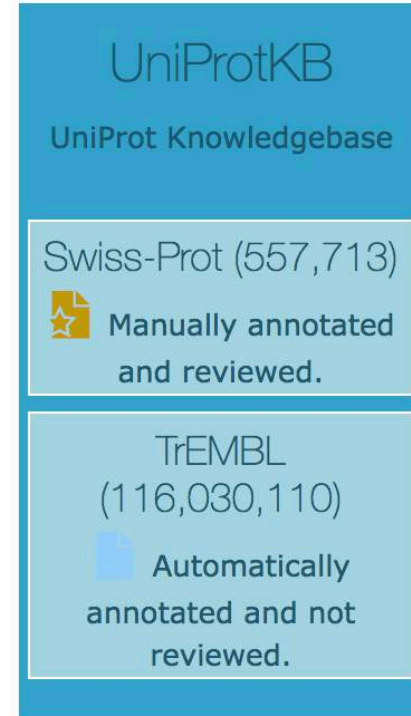
Doğrulama veri kümesi kullanılarak

$$W(m) = \frac{R_m^4}{R_{BLAST-kNN}^4 + R_{SPMap}^4 + R_{PEPSTATS-SVM}^4}$$

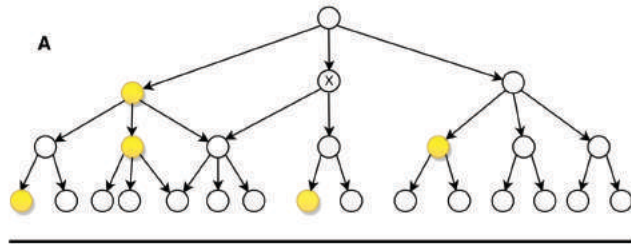
Veri

Universal Protein Resource Knowledge Base (UniProtKB)

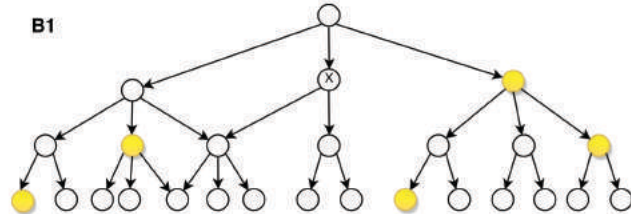
- Proteinlerin dizi ve işlevsel bilgileri
- Küratörler literatür taraması yapıp bilgileri toplar
- Araştırmacıların erişimine sunulur



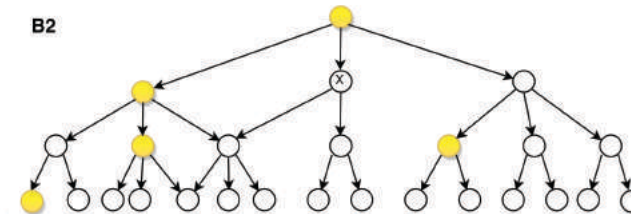
Pozitif ve Negatif veri kümesi oluturulması



(A) X GO terimi için pozitif veri kümesi oluşturulması
P1 diye hipotetik bir protein ve işlev (terim) atamaları sarı ile gösterilmiş olsun
X'in çocuk düğümündeki terim ile işlev ataması yapılmış olduğu için P1 X'in pozitif veri kümesinde yer alır



(B1) P2 X'in ebeveyniyle anlamlandırılmadığı için P2 X'in negatif veri kümesine dahil edilir



(B2) P3 X'in ebeveyniyle anlamlandırılmış ama aynı anda kardeşi ile de anlamlandırılmış (aynı düzeyde farklılaşmışlar). P3 X'in negatif veri kümesine dahil edilir

Veri

- Verinin eskisi gibi az değil, ama kirli
- Sayılar büyük
 - <https://www.uniprot.org/>
116,030,110 dizisi bilinen ama anlamlandırılmamış protein
 - <https://pubchem.ncbi.nlm.nih.gov/#>
Compounds: 96,398,651
Substances: 246,968,148
BioAssays: 1,251,890
BioActivities: 236,710,461
Protein Targets: 10,854
- Bu verileri anlamlandırmak ve fazlalıkları (redundancy) gidermek için bir çok otomatik yöntem mevcut

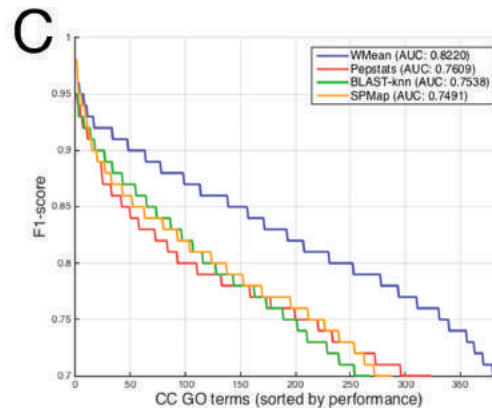
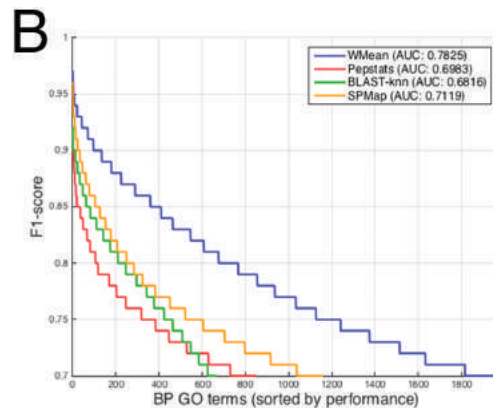
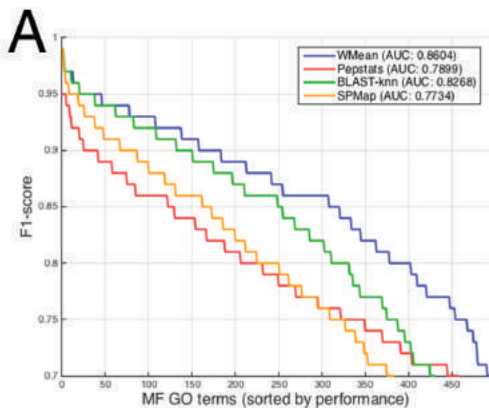
Doğrulama, Başarım,
Karşılaştırma,

Başarım ve Karşılaştırma

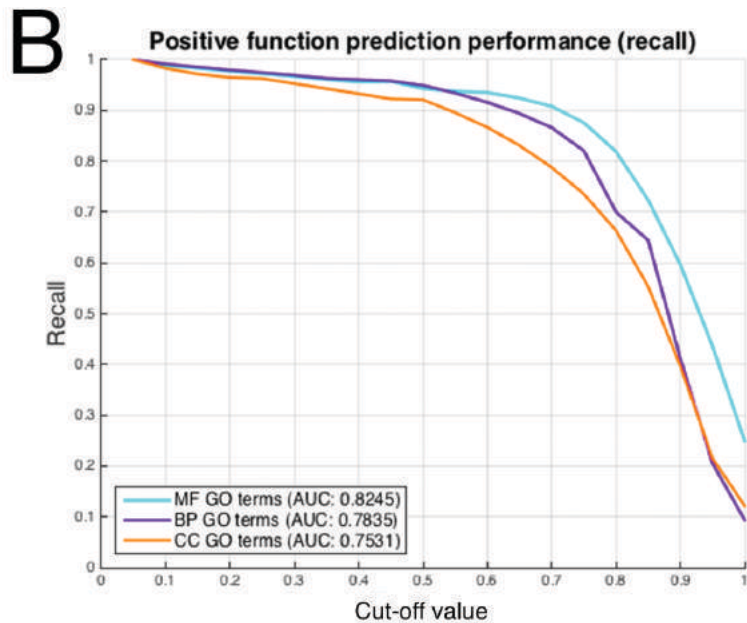
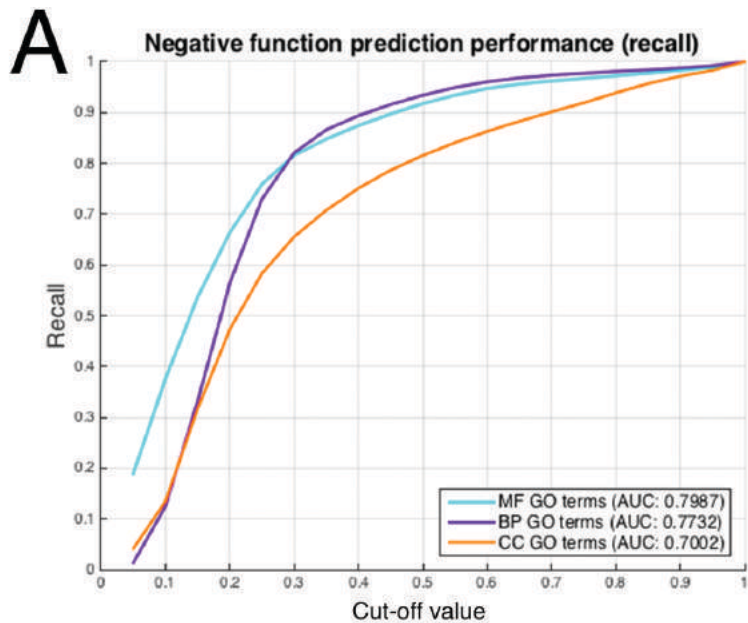
- Doğrulama
- Bağımsız test kümeleri
- Diğer benzer sistemlerle karşılaştırma
- Yarışmalar

- F1-skoru
- kesinlik (precision) ve duyarlılık (recall)

UniGOPred



UniGOPred



ECPred-Doğrulama sonuçları

Eğitimde kullanılmamış %20 veri

EC Level	F1-score	Recall	Precision
Level 0	0.96	0.96	0.96
Level 1	0.96	0.96	0.96
Level 2	0.98	0.97	0.99
Level 3	0.99	0.98	0.99
Level 4	0.99	0.99	0.99

Bağımsız test kümeleri

- Veritabanının iki sürümü arasında anlamlandırılan proteinler
 - Temporal Hold-out Dataset Test
- Başka sistemlerde kullanılmış olan veri kümeleri
 - Eğitim kümesinde olan proteinler kullanılmadı

Diğer benzer sistemlerle karşılaştırma

- Mutlaka yapılmalı
 - Saygın dergiler istiyor
- Verilen kodları veya web tabanlı servisleri kullanarak
- Veya kullandıkları test veri kümelerini kullanarak

Bağımsız test-zaman içinde değişen veriler

0. düzey-enzim and enzim değil

Method	F1-score	Recall	Precision
ProtFun	0.79	0.87	0.72
EzyPred	0.15	0.13	0.16
EFICAz	0.42	0.30	0.69
DEEPre	0.53	0.43	0.68
ECPred	0.83	0.97	0.73

1. düzey-ana sınıf

Method	F1-score	Recall	Precision
ProtFun	0.12	0.10	0.15
EzyPred	0.15	0.13	0.16
EFICAz	0.42	0.30	0.69
DEEPre	0.50	0.40	0.67
ECPred	0.48	0.43	0.54

2. düzey-alt sınıf

Method	F1-score	Recall	Precision
EzyPred	0.11	0.10	0.13
EFICAz	0.11	0.07	0.33
DEEPre	0.11	0.25	0.07
ECPred	0.26	0.20	0.35

3. düzey-alt alt sınıf

Method	F1-score	Recall	Precision
EFICAz	0.00	0.00	0.00
DEEPre	0.05	0.03	0.14
ECPred	0.22	0.17	0.31

PFAM bölge (domain) içermeyen

0. düzey-enzim and enzim değil

Methods	F1-score	Recall	Precision
EzyPred	0.54	0.54	0.54
EFICAz	0.37	0.23	1.00
DEEPre	0.60	0.4	0.85
ECPred	0.85	0.82	0.89

1. düzey-ana sınıf

Methods	F1-score	Recall	Precision
EzyPred	0.42	0.39	0.46
EFICAz	0.33	0.20	1.00
DEEPre	0.52	0.38	0.82
ECPred	0.73	0.63	0.86

2. düzey-alt sınıf

Methods	F1-score	Recall	Precision
EzyPred	0.30	0.26	0.36
EFICAz	0.33	0.20	1.00
DEEPre	0.40	0.27	0.77
ECPred	0.60	0.47	0.82

3. düzey-alt alt sınıf

Methods	F-score	Recall	Precision
EFICAz	0.33	0.20	1.00
DEEPre	0.33	0.22	0.73
ECPred	0.58	0.45	0.81

4. düzey-substrat sınıfı

Methods	F-score	Recall	Precision
EFICAz	0.33	0.20	1.00
DEEPre	0.33	0.22	0.73
ECPred	0.39	0.26	0.74

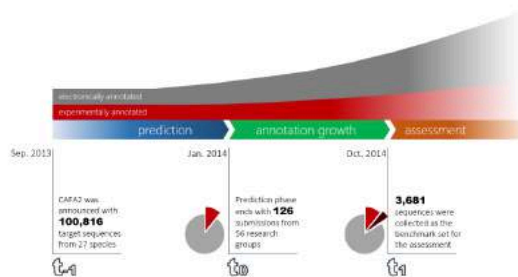
CAFA <http://biofunctionprediction.org/cafa/>

Function Special Interest Group

About Meetings CAFA Publications

The CAFA Challenge:

The problem: There are many proteins in the databases for which the sequence is known, but the function is not. The gap between what we know and what we do not know is growing. A major challenge in the field of bioinformatics is to predict the function of a protein from its sequence or structure. At the same time, how can we judge how well these function prediction algorithms are performing?



The solution: The Critical Assessment of protein Function Annotation algorithms (CAFA) is an experiment designed to provide a large-scale assessment of computational methods dedicated to predicting protein function, using a time challenge. Briefly, CAFA organizers provide a large number of protein sequences. The predictors then predict the function of these proteins by associating them with Gene Ontology terms or Human Phenotype Ontology terms (Blue "prediction" section of timeline). Following the prediction deadline, we wait for several months. During that time, some proteins whose function were unknown experimentally have received experimental verification (Green "annotation growth" section of timeline). Those proteins constitute the benchmark, against which the methods are tested (Orange "assessment" portion of timeline). You can read about CAFA 1 [here](#) and in the paper published in *Nature Methods*, and you can read about CAFA 2 [here](#).

CAFA π

Job Opportunity

There is an opportunity for a postdoc / research scientist to run CAFA3. For position details see: <https://careers.iscb.org>

Submission Deadline

000 : 00 : 00 : 00
Day : Hrs : Min : Sec

CAFA PI open for registration!

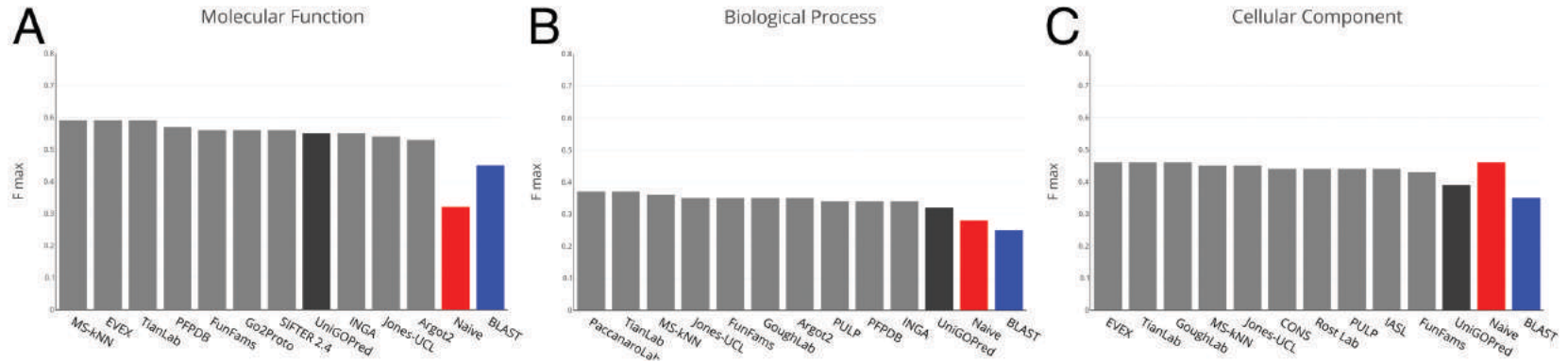
Click to
Participate

Target release date: December 1, 2017.

Predictions deadline: April 20, 2018.

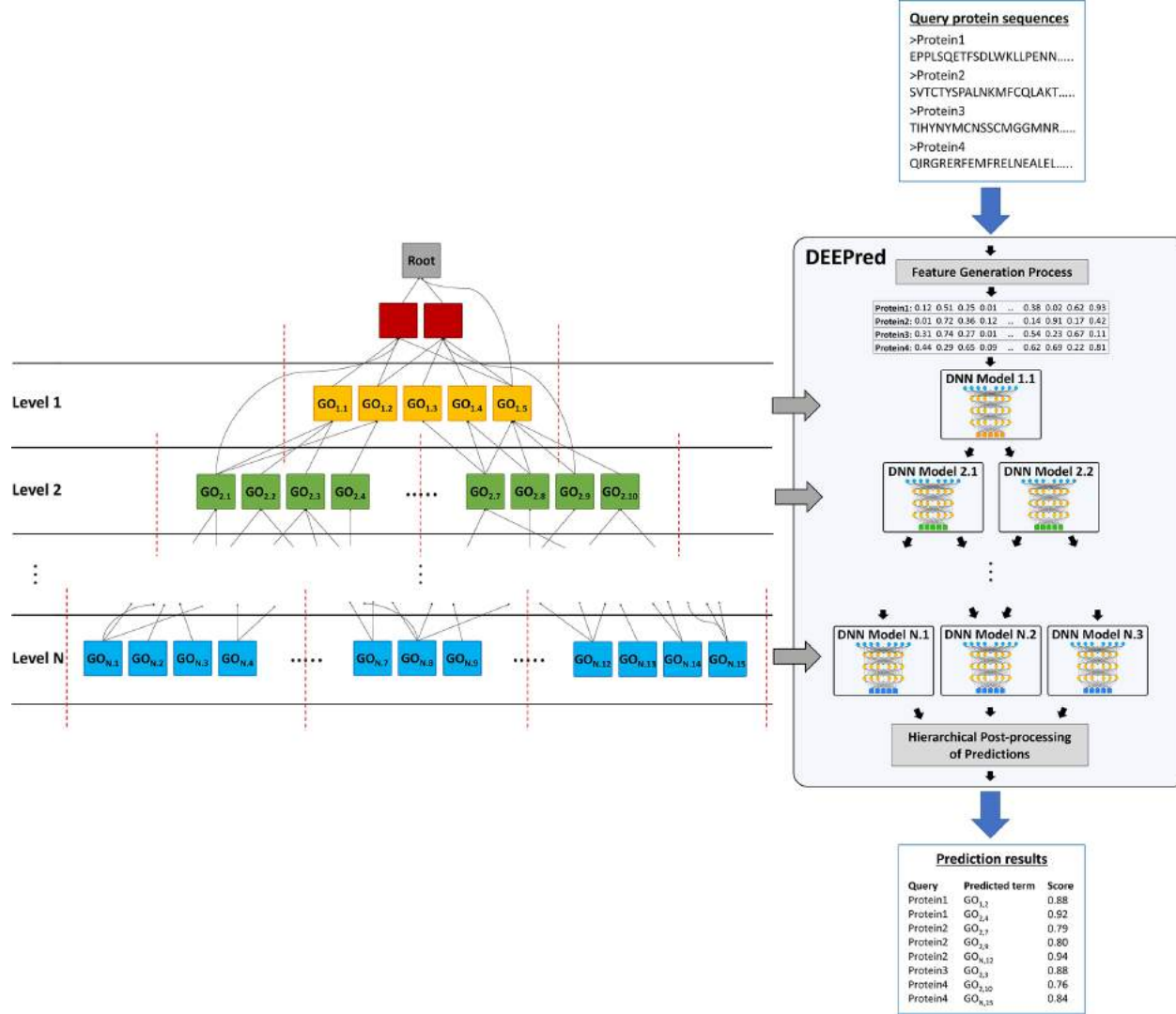
Initial Evaluation: July 2018.

CAFA sonuçları ile karşılaştırma



Çoklu sınıflandırıcı

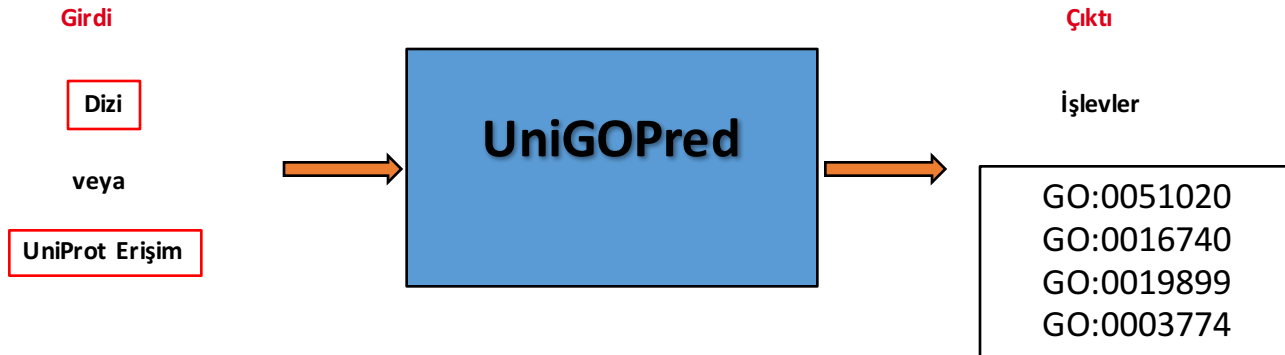
- Derin öğrenme ile
- Ama tüm terimler değil-en fazla 5-7 terim için bir sınıflandırıcı



Servisler

UniGOPred

- **UniGOPred** kapsamlı bir protein anlamlandırma sistemi-iki bileşen:
 - dizi tabanlı GO'ya dayalı ileri öngörü aracı
 - UniProtKB proteinleri için öngörölmüş GO terimlerinin veritabanı
- Hem sınıflandırma hem de transfer yaklaşımı ve üç yöntemin birleşimi
- Yeni bir veri hazırlama yaklaşımı
- <http://cansyl.metu.edu.tr/unigopred.html>



<http://cansyl.metu.edu.tr/unigopred.html>



GENE ONTOLOGY PREDICTION BY COMBINING CLASSIFIERS

[Home](#)

[About](#)

[Help](#)

[Contact](#)

UniGOPred is an automated protein function prediction tool based on Gene Ontology (GO) terms and a database of GO term predictions for UniProtKB. You can get MF predictions for all **UniProtKB** sequences. BP and CC GO term predictions are also available for all sequences of **Model Organisms**. You can submit sequence(s) (up to 10) or UniProt accessions (up to 100) to get your predictions! [Here](#) is the valid input formats that UniGOPred accepts. Prediction scores for the trained GO terms will be sent to your e-mail address. Further information about UniGOPred and trained GO terms is available in [About](#) page.

Query Type:

☒ Pre-computed Database Predictions

☐ Sequence

☒ Molecular Function ☐ Biological Process ☐ Cellular Component

Please enter your e-mail address :

Please enter UniProt Accession(s) :

Accession(s)

Get
Predictions

<http://cansyl.metu.edu.tr/ECPred.html>

ECPred

Prediction of enzymatic properties of protein sequences based on the EC Nomenclature

The automated prediction of the enzymatic functions of uncharacterized proteins is a crucial topic in bioinformatics. Although several methods and tools have been proposed to classify enzymes, most of these studies were limited to specific functional classes and levels of the Enzyme Commission (EC) number hierarchy. Besides, most of the previous methods incorporated only a single input feature type, which limits the applicability to the wide functional space. Here, we proposed a novel enzymatic function prediction tool, ECPred, based on ensemble of machine learning classifiers. In ECPred, each EC number constituted an individual class and therefore, had an independent learning model. Enzyme vs. non-enzyme classification is incorporated into ECPred along with a Hierarchical prediction approach exploiting the tree structure of the EC nomenclature. ECPred provides predictions for 658 EC numbers in total including 6 main classes, 65 subclass classes, 163 sub-subclass classes and 634 substrate classes. The proposed method is tested and compared with the state-of-the-art enzyme function prediction tools by using independent temporal hold-out and no-Pfam datasets constructed during this study. ECPred is presented both as a stand-alone and a web-based tool to provide probabilistic enzymatic function predictions (at all five levels of EC) for uncharacterized protein sequences. Also, the datasets of this study will be a valuable resource for future benchmarking studies. ECPred is available for download, together with all of the datasets used in this study, at: <https://github.com/cansyl/ECPred>.

To download ECPred stand-alone tool: [Click Here](#)

ECPred github link: github.com/cansyl/ECPred

Here is the valid input formats that ECPred accepts. Prediction for the given sequence will be sent to your e-mail address.

Please enter your e-mail address :

Please enter your sequence here!

Benzer çalışmalarımız

- İşlev Öngörüsü
 - ECPred <http://cansyl.metu.edu.tr/ECPred.html>
 - DEEPred
- Sanal Tarama
 - ConvNet

Yeni Yayınlar

- A.S. Rıfaioglu, H. Atas, M.J. Martin, R. Cetin Atalay, V. Atalay, T. Dogan, "Recent Applications of Deep Learning and Machine Intelligence on In-silico Drug Discovery: Methods, Tools and Databases", [Briefings in Bioinformatics](#) to appear, 2018. 
- A.S. Rıfaioglu, V. Atalay, M.J. Martin, R. Cetin Atalay, T. Dogan, "Drug-Target Interaction Prediction with Deep Convolutional Neural Networks Using Compound Images", Oral presentation, [Machine Learning in Computational and Systems Biology \(MLCSB\) COSI at ISMB 2018](#). 
- T. Dogan, A.S. Rıfaioglu, R. Saidi, M.J. Martin, V. Atalay, R. Cetin Atalay, "Automated Negative Gene Ontology Based Functional Predictions for Proteins with UniGOPred", Oral presentation, [Function-COSI at ISMB 2018](#). 
- H.F. Danaci, R. Cetin Atalay, V. Atalay, "[EClerize: A customized force-directed graph drawing algorithm for biological graphs with EC attributes](#)", *Int. Journal of Bioinformatics and Computational Biology*, to appear, 2018.
- A.S. Rıfaioglu, T. Doğan, Ö.S. Saraç, T. Ersahin, R. Saidi, M.V. Atalay, M.J. Martin, R. Cetin Atalay, "[Large-scale automated function prediction of protein sequences and an experimental case study validation on PTEN transcript variants](#)", *Proteins: Structure, Function, and Bioinformatics*, 2018;86:135–151. <https://doi.org/10.1002/prot.25416> 
- A.S. Rıfaioglu, T. Dogan, M.J. Martin, R. Cetin Atalay, V. Atalay, "[Multi-task Deep Neural Networks in Automated Protein Function Prediction](#)", preprint, arXiv:1705.04802.
- A.S. Rıfaioglu, M.J. Martin, R. Cetin Atalay, V. Atalay, T. Dogan, "Investigation of Multi-task Deep Neural Networks in Automated Protein Function Prediction", Oral Presentation at *Function COSI ISMB/ECCB 2017: 25th Annual International Conference on Intelligent Systems for Molecular Biology*, Prague, Czech Republic, July 2017.
- A. Dalkiran, A.S. Rıfaioglu, T. Dogan, V. Atalay, M.J. Martin, R. Cetin Atalay, "Prediction of Enzymatic Properties of Protein Sequences Based on the EC Nomenclature", Poster Presentation at *ISMB/ECCB 2017: 25th Annual International Conference on Intelligent Systems for Molecular Biology*, Prague, Czech Republic, July 2017.
- A.S. Rıfaioglu, M.J. Martin, R. Cetin Atalay, V. Atalay, T. Dogan "Investigation of Multi-task Deep Neural Networks in Automated Protein Function Prediction", Oral Presentation at *HIBIT 2017: The International Symposium on Health Informatics and Bioinformatics*, METU Northern Cyprus Campus, June 28 – 30, 2017.
- H.F. Danaci, R. Cetin Atalay, V. Atalay, "EClerize: A Customized Force-Directed Layout Algorithm for Biological Networks with EC Attributes", Oral Presentation at *GLBIO 2017: Great Lakes Bioinformatics Conference*, Chicago, USA, May 15-17, 2017.
- A.S. Rıfaioglu, T. Dogan, Ö.S. Sarac, R. Saidi, V. Atalay, M.J. Martin, R. Cetin Atalay, "UniGOPred: A Large Scale Automated GO Term Annotation System for UniProtKB", Poster presentation at *GLBIO 2017: Great Lakes Bioinformatics Conference*, Chicago, USA, May 15-17, 2017 (Best Poster Award).

- **Destek:** Newton-Kâtip Çelebi Fonu İkili İşbirliği Programı 2016 çağrısı
British Council, UK and TÜBİTAK, Türkiye
- **Takvim** : 1 Kasım 2017 – 31 Ekim 2019
- **Proje Başlığı:**
 - "Derin Öğrenme Teknikleri Ve Ağ Analizi Yöntemleriyle Hazırlanmış Kapsamlı Biyomedikal İlişkiler Kaynağı"
 - "Comprehensive Resource of Biomedical Relations with Deep Learning and Network Representations"

<u>TR</u>	<u>UK</u>
Mehmet Volkan Atalay (PI)	Maria Martin (PI)
Rengul Atalay	Rabie Saidi
Tunca Doğan	Andrew Nightingale
Ahmet Rifaioğlu	Vladimir Volykin
Heval Ataş	Vishal Joshi
Alperen Dalkıran	
Nurcan Tunçbağ (akademik danışman)	
Tuğba Süzek (akademik danışman)	

Destekler

- TÜBİTAK EEEAG 116E930 Newton-Katip Çelebi British Council-TÜBİTAK İkili İşbirliği Projesi
- Kalkınma Bakanlığı





HIBIT2018

[Home](#)

[Registration](#)

[Submission](#)

[Key Dates](#)

[Main Program](#)

[Travel & Lodging](#)

[Venue](#)

[Committees](#)

[Contact Us](#)

Welcome to

The 11th **HIBIT** Conference

International Symposium on Health Informatics
and Bioinformatics

October 25-27, 2018 - Antalya

Ekibimiz



Resimde olmayanlar

Ö. Sinan Saraç

Ayşe Gül Yaman

Alperen Dalkıran

EBI

Maria Martin

Rabie Saidi

Heval Ataş

Rengül
Çetin-Atalay

Volkan Atalay

Ahmet
Rıfaioğlu

Tunca Doğan